Post Editing Guidelines
For
GALE Machine Translation Evaluation

Version 2.0.7
2006-04-21

National Institute of Standards and Technology
Information Access Division / Speech Group
And Linguistic Data Consortium

http://www.nist.gov/speech/tests/GALE

1 Introduction

Machine Translation (MT) is a technology that translates language data from a foreign language into a target language. The source language for the GALE program will be either Arabic or Chinese, and the target language will English. Input data will be in the form of either audio or text, with the output always being text.

The GALE program will evaluate MT in terms of the quality of the system translations. This will be accomplished by measuring the edit distance between a system output and a gold standard reference. The term "edit-distance" refers to the minimum number of edits (modifications) that someone *needs* to make to the output of a machine translation system such that the resulting text completely captures the meaning of the gold standard reference using English that is as understandable as the reference.

This document serves as a reference for those tasked with post-editing machine translation output for the GALE evaluations. Post editors work as independent contractors. Each post editor will download a set of system output and gold standard reference files along with a specialized annotation tool to his or her home computer and will edit the system output translations according to the guidelines outlined in the sections below.

Data is divided into units of comparable size and difficulty, called "kits." Post editors are supplied with a workflow management interface that permits them to check kits out and in. Supervisors are alerted to any new kit submissions and check the kits for quality. Kits that do not comply with the guidelines outlined in this document are returned to the post editors with feedback.

Section 2 of this document describes the role of the editor. Section 3 is organized so that each main point addresses a single high level rule, followed with examples and in rare occasions, exceptions. A short summary of each high level rule can be found in Appendix A.

Post Editors should read through the entire document before beginning the editing process, and refer to it as necessary, but Appendix A was designed to serve as a one-page crib sheet.

A training session is provided for each post editor, during which editing interface is explained and any questions about the process or the project goals are answered.

2 Post Editor Role

The job of the MT post editors may be summarized as follows:

> **"Make the MT output have the *correct meaning,* using *understandable* English, in as *few edits* as possible."**

The Post Editor should keep in mind that:

- There are equally-accurate alternative ways to edit the machine translation output, all being considered correct, some using fewer edits than others.

- The reference translation is created by humans and may on occasion contain what appears to be an error. In such cases the editor should treat the reference as being absolutely correct (editors are not to "repair" the reference).
- The resulting edited version should have the same meaning as the reference translation; the edited version should not *add* any information that is *not* present in the reference, and the edited version should not *omit* any information that *is* in the reference.
- The editing occurs on a segment-to-segment basis to make it more manageable for the post editors, but it is more correct to view these rules as applying to the document as a whole. In particular, translation systems and your edited versions are free to rearrange words and phrases over long distances (translation is not a word-by-word or phrase-by-phrase mapping process).
- Minimize the number of edits whenever that can be done without compromising the meaning and without making the edited version hard to understand.
- Avoid wholesale copying of the reference.

2.1 Understanding Post Edit Alignment

As stated above:

> "…editing occurs on a segment-to-segment basis to make it more manageable for the post editors, but it is more correct to view these rules as applying to the document as a whole. …"

There are many variations of valid translations. It is perfectly acceptable for the machine translation output to put some phrases in different segments than where the reference has them. For this editing task, the system translations have been aligned to the reference translations automatically. There will undoubtedly be instances where the system translation segment being edited contains information which is in the preceding (or following) reference segment. This is okay, and the editor **must not** move the information just to make it align with the reference data. Table 1 shows an example of differing alignments that does NOT require any editing.

Reference Data	System Translation to Edit
The President signed the Transportation bill on Tuesday morning, and flew to Camp David later that evening.	Being overdue for an extended vacation, the President flew to Camp David on Tuesday evening.
He was overdue for an extended vacation.	On Tuesday morning he signed the Transportation bill.

3 Editing Rules

This section outlines the rules for each of the guidelines identified in Appendix A. The post editor must strictly adhere to these rules.

3.1 *Capturing the Meaning of the Reference*

> ***Rule #1***:
> Make the MT output have the **same meaning** as the reference
> human translation. No more and no less.

Rule #1 represents the primary rule of post editing. Under no circumstance should the post editor sacrifice capturing the meaning of the reference translation in order to better meet the requirements of another rule. The English language allows the same information to be presented in several ways. The post editor should focus on capturing the meaning in the reference translation, not necessarily creating a word-for-word parroting of the reference text.

It is important to capture the complete meaning. At times the MT output will be missing information that will need to be added while other times the MT output will include extra information that must be removed. Consider the reference to be *always* correct.

3.1.1 Synonyms

If the meaning is correct, synonyms are completely acceptable.

3.1.2 Acronyms

Acronyms **can** be substituted for the expanded written form, and visa versa.

3.1.3 Symbols

Symbols such as the dollar sign '$', percent symbol '%' and the pound sign '#' are equivalent to the written form and should not be edited.

The following example is fine as is, and need not be edited:

Reference: The company stock rose by 6 percent to $22.50.
MT output: The company stock rose by six% to twenty-two dollars and fifty cents.

3.1.4 Numbers

Numbers may be either spelled out as in "one-hundred and one" or written using numbers as in "101". Do not change either form to the other. Do not insert or remove hyphens in forms like "twenty-two".

3.1.5 Abbreviations

Abbreviations are completely acceptable. "Mr." for "Mister" and "lbs" for "pounds" are just two examples.

3.1.6 Contractions

Contractions are perfectly legitimate alternatives to the full spelling, that is "can't" is equivalent to "cannot".

3.1.7 Phrasal Ordering

The order of phrases only needs to be changed to correct the meaning or to follow other guidelines in this document. For instance, the following is fine as is, and need not be edited:

Reference: Since the light was red, he stopped his car.
MT output: He stopped his car, since the light was red.

But the following example requires a reordering of the phrases in order to capture the meaning of the reference translation:

Reference: The protestors rallied against the government to change the outdated law.
MT output: The protestors will modify the outdated law to rally against the government.
Edited MT: The protestors rallied against the government to modify the outdated law.

Phrasal ordering may improve **English Understandability**, see section 3.5.

3.1.8 Parts of Speech

It is important that post editors carefully consider differing parts-of-speech to be sure that the complete and correct meaning of the reference data is captured in the MT output and made readily understandable.

3.1.8.1 Verb Tense

The MT output must have the same verb tense as the reference when this makes a difference in the meaning. Verb tense may affect English understandability, see section 3.5.

3.1.8.2 Prepositions

Prepositions may need to be changed, inserted or deleted, in order to correct the meaning.

Reference: The rebels looked at the hills *beyond* the road after the smoke cleared.
MT output: When the rebels looked up the hills *under* the road the smoke cleared.
Edited MT: The rebels looked at the hills beyond the road when the smoke cleared.

In this example the prepositions "up" and "under" must be changed.

3.1.8.3 Adverbs and Adjectives

At times adverbial and adjectival constructions will need to be changed or added to relate the correct meaning. If the reference used the phrases "*lofty* goals" and "*dangerous* job", the meanings conveyed by both lofty and dangerous must be included in the MT output.

In some cases, the need to make an addition or change will be a judgment call by the post editor.

3.1.8.4 Determiners

Determiners are sometimes required to convey the complete meaning, as in:

Reference: …representing *the* different historical periods of Egypt.
MT output: …representing different historical periods of Egypt.

Subtle, but the determiner "*the*" is required to communicate the fact that **all** the historical periods are represented.

3.1.8.5 Pronouns

Pronouns should be correct and may be used in place of their corresponding noun if the meaning is correct.

3.1.9 Extra Information in MT Output

If it is the case that the MT output contains <u>*information*</u> that is not in the reference translation, delete it. In the following example "evening" should be deleted from the MT output:

Reference: The bomb went off on Tuesday.
MT output: The bomb exploded on Tuesday *evening*.
Edited output: The bomb exploded on Tuesday.

> **COMMENT**: This was a simple example. The post editor was able to make a decision based on a single segment. It is the case however, that at times the post editor will have to consider segments *before* and *after* the one being edited in order to correctly determine if the information is indeed extra rather than simply moved from a different sentence/segment.

The structure of the English language is very different from some languages. English might convey information by inflecting verbs or by the form of a noun or pronoun where other languages convey the same meaning by means of more verbose syntax.

It is okay if the MT output has the *same* meaning as the reference but conveys that meaning in a more wordy fashion. For example it would be considered correct if the MT output had "five head of cattle", where the reference had something like "five cattle". However, there are many wordy phrases that do convey some meaning. Editing should never compromise the meaning. So, phrases such as "in the case of", "for example", "a number of", and "on behalf of" should be added or removed to match the meaning of the reference whenever they make any difference in the meaning. ***When required to make an addition, use the shortest phrase (or single word) that conveys the same meaning***.

3.1.10 Information Missing from MT Output

This is the counterpart of the preceding rule. If some of the <u>*meaning*</u> of the reference is missing from the MT output, it must be added. The MT output is to be a complete translation.

Reference: The interim Iraqi government will hold its first election on January 30. The US has deployed nearly 20,000 additional troops to ensure the security of the *Iraqi* election.

MT output: The Iraqi interim government election is scheduled to be held on January 30. The United States has deployed some 20,000 troops to protect the security of the election.

Edited MT: The Iraqi interim government*'s* *first* election is scheduled to be held on January 30. The United States has deployed some 20,000 troops to protect the security of the election.

Note:

 (1) The fact that this was the "first" election was added.

 (2) The last occurrence of "Iraqi" in the reference was redundant and not added.

Since post editors are editing the translation as a whole, not by each individual word or phrase, it is okay if the MT omits duplicated or redundant words and phrases when the meaning is *unambiguously complete and correct* (as was the case with the last occurrence of "Iraqi" in the preceding example)

In the following example, the MT is acceptable **only if** because of other context we know that I fly in an airplane, rather than a helicopter or whatever.

Reference: I fly in an airplane from Shanghai to Beijing.
MT output: I fly from Shanghai to Beijing.

3.2 Making the MT Understandable

+---+

***Rule #2**:*
Make the MT output be *as* **understandable** as the reference.

+---+

Post editors should make the MT output as understandable as the reference translation, but must not try to make it more understandable.

> *Note: While every effort is being made to construct reference data that is accurate and fluent English, there is always the possibility of human error (spelling, punctuation, etc.) Always treat the reference translation as correct.*

3.2.1 Reference Ambiguity

At times the reference data will be fluent, but not understandable. Consider the following example:

Reference: Journalists and the media were only allowed to tell the people what the data or government dictated that they say.

MT output: Journalists and men information, only to convey to the people the dictates of statements by the government.

There is not a clear distinction in the *reference* between "journalists" and "the media". "The data" does not have a clear distinction either. Here is what three editors did for this example:

Post-editor1: Journalists and the media were only allowed to convey to the people the dictates of the data or government.

Post-editor2: All that was available to journalists or media was to convey to the people the statements dictated by the data or government.

Post-editor3: Journalists and media were only allowed to convey to the people what the data or government dictated they say.

Note that all three editors were scrupulous about the meaning. Note ***carefully*** that the editors did not make the MT output *more* understandable than the reference, because doing so might have involved some compromise about the meaning of the reference.

3.3 Minimizing the Number of Edits

Rule #3:

Capture the correct meaning in **as few edits as possible** using understandable English.

In case of perceived conflicts among the parts of this rule: correcting the meaning and making the output as understandable as the reference are more important than minimizing the number of edits.

Because minimizing the number of edits is a goal, when content from the reference is *missing* from the MT output, try to *insert* that missing reference content as the shortest replacement that you can think of with exactly the right meaning.

Post editors should limit the use of editorial license in altering the existing MT output to a minimum. Do not delete or change perfectly acceptable words/phrases/punctuation that exist in the MT output to some substitute just because ***you*** have a preference for something else. Keep in mind that minimizing the number of edits is a goal. One way to think about minimizing edits is that the resulting edited version should preserve (completely unmodified) as much of the original MT output as possible without compromising the other guidelines; note that modifying the form of a word (for example, changing "flyer" to "flyers" or changing "radiates" to "radiating") is no more desirable than substituting a completely different word.

Perhaps the most interesting challenge about doing your editing task is to be so clever that you find a way to fix the meaning with the minimum number of edits. This is very important goal.

3.3.1 Spelling

Do not fix spelling variants in spelling such as "the colour grey" that reflect British usage.

3.3.2 Proper Names

Spelling variants are acceptable.

Sometimes alternate phrases will convey the same meaning as a proper name, for example: "North Korea" and "northern Korea" and "Democratic People's Republic of Korea" and "People's Democratic Republic of Korea" represent acceptable alternatives.

3.3.3 Dates

Dates must be correct. The date format in the MT output does not have to match the reference, but it must be correct and unambiguous. Dates may be represented in one of many numeral formats. For example, the following are all equivalent:

- 02/14/2006
- 2-14-2006
- 2.14.06

The European equivalent is acceptable when no confusion would arise as in 14/02/06.

Dates may be spelled out in part or in whole (abbreviations are completely acceptable). For example, the following are all equivalent:

- February 14th, 2006
- Feb. 14 '06
- Feb/14/2006
- 14/Feb/2006

In the following example no editing is required:

Reference: The bomb exploded on 7/21/1989.
MT output: The bomb detonated on the 21st of July, 1989.

3.3.4 Decimal Points

It may be the case that a comma is used in place of the decimal point. Post editors should correct this "European" format to the American format.

3.4 Punctuation and Capitalization

> ***Rule #4***:
> Punctuation must be understandable, and sentence-like units must have sentence-ending punctuation and proper capitalization. Do not insert, delete, or change punctuation merely to follow traditional optional rules about what is "proper".

If the punctuation in the MT output is acceptable, leave it unchanged.

3.4.1 Sentence Ending Punctuation

Insert missing sentence-ending punctuation, so that each sentence-like unit ends with sentence-ending punctuation as in the reference translation. Choose from a period '.', a question mark '?', or an exclamation mark '!'.

If the MT output is acceptable do not split or merge sentences just to match the reference.

Headlines, book titles, and so forth do not have to be punctuated as a sentence.

3.4.2 Quotation Marks

Single and double quotation marks must be properly paired and quote the correct material. Note, that sometimes the quotation marks will span more than the segment currently being edited and may require looking ahead to see where the quote ends.

3.4.3 Change/Insert/Delete Punctuation

Punctuation should be corrected as needed to repair or disambiguate the meaning or to make the output understandable. Do not add/delete/change sentence-internal punctuation unless it is necessary for understandability or makes a difference in the meaning. The edited version does not have to match all the semicolons, colons, dashes, ellipsis dots (...) and commas in the reference, nor does the edited version have to follow traditional "rules" about what is correct sentence-internal punctuation.

3.4.4 Capitalization

It is important that the edited machine translation output have proper capitalization, both for proper nouns, and for proper sentence construction. Names of persons and places must be capitalized. Capitalization of job titles like "National Police Chief" is optional but must be consistent within a sentence. In the case of headlines, book titles, and so forth, there is no standard capitalization—only make sure that the first word of them is capitalized and that proper names are capitalized.

3.5 Odd but Understandable English

At times, the translation will be clearly understandable, yet odd. At times, what is odd may be the order of phrases, reflecting the structure of the foreign language in the source material. Here are some examples where the order of phrases reflects the structure of a foreign language that would normally put the topic first, when English would normally put the subject of the verb first. As long as the MT output is readily understandable (as in these examples, if you find them readily understandable), don't rearrange the phrases just to improve the fluency.

Examples:

MT: This book I have read.	**Reference**: I have read this book.
MT: In the yard is parked a car.	**Reference**: A car is parked in the yard.
MT: Swimming, I am the best.	**Reference**: I am the best at swimming.
MT: A dog have I.	**Reference**: I have a dog.

Post editors will, however, usually need to insert missing verb tenses/inflections when the MT is not reasonably fluent English, in order to capture the meaning.

4 Notes Regarding Conversational Speech

Some of the data being edited will be spontaneous speech and will have a conversational style. This section addresses some of the nuances associated with this type of data.

4.1 Filled Pauses

Filled pauses are occurrences of lexical tokens such as {uh, um, er, huh, uh-uh, …}. These tokens may appear in either the references or the system translations. These tokens need not be preserved. They should neither be removed nor added during post editing.

4.2 Repairs

A "repair" occurs when a speaker corrects their speech as in: "John told me on **Tuesday, I mean on** Wednesday that the due date has changed." "Tuesday, I mean on" is a repair that should be present in the edited MT.

4.3 Completely Incoherent MT System Translations

At times the MT system translation may appear as complete garbage. Avoid the temptation of erasing the MT system translation and copying in the reference translation, rather the post editor should try in earnest to salvage as much of the MT system translation as possible. Consider the following example:

Reference: The interim Iraqi government will hold its first election on January 30. The US has deployed nearly 20,000 additional troops to ensure the security of the *Iraqi* election.
MT output: Middle ground was found on the first day of January despite the US ability to send more aid for travel arrangements.

First DELETE: ~~Middle ground was found on t~~he first ~~day of~~ January ~~despite t~~he US ~~ability~~ to sen~~d~~ more ~~aid~~ for ~~travel arrangements~~.

Second INSERT: The interim Iraqi government will hold its first election on January 30. The US has sent 20,000 more troops to ensure the security for the Iraqi election.

5 Appendix A

Appendix A contains the guidelines for editing the machine translation output (MT output). This is the set of rules each post editor should keep near them while they edit until they are committed to memory. ***Minimizing the number of edits is a goal***, ***but your primary goal is to make the meaning correct while making the edited version understandable.***

Make the MT output have the *correct meaning,* using *understandable* English, in as *few edits* as possible.

1. Make the MT output have **the same meaning** as the reference human translation. No more and no less.

2. Make the MT output be as **understandable** as the reference.

3. Capture the meaning in **as few edits as possible** using understandable English. If words/phrases/punctuation in the MT output are completely acceptable, use them (unmodified) rather than substituting something new and different.

4. Punctuation must be understandable, and sentence-like units must have sentence-ending punctuation and proper capitalization. Do not insert, delete, or change punctuation merely to follow traditional optional rules about what is "proper."